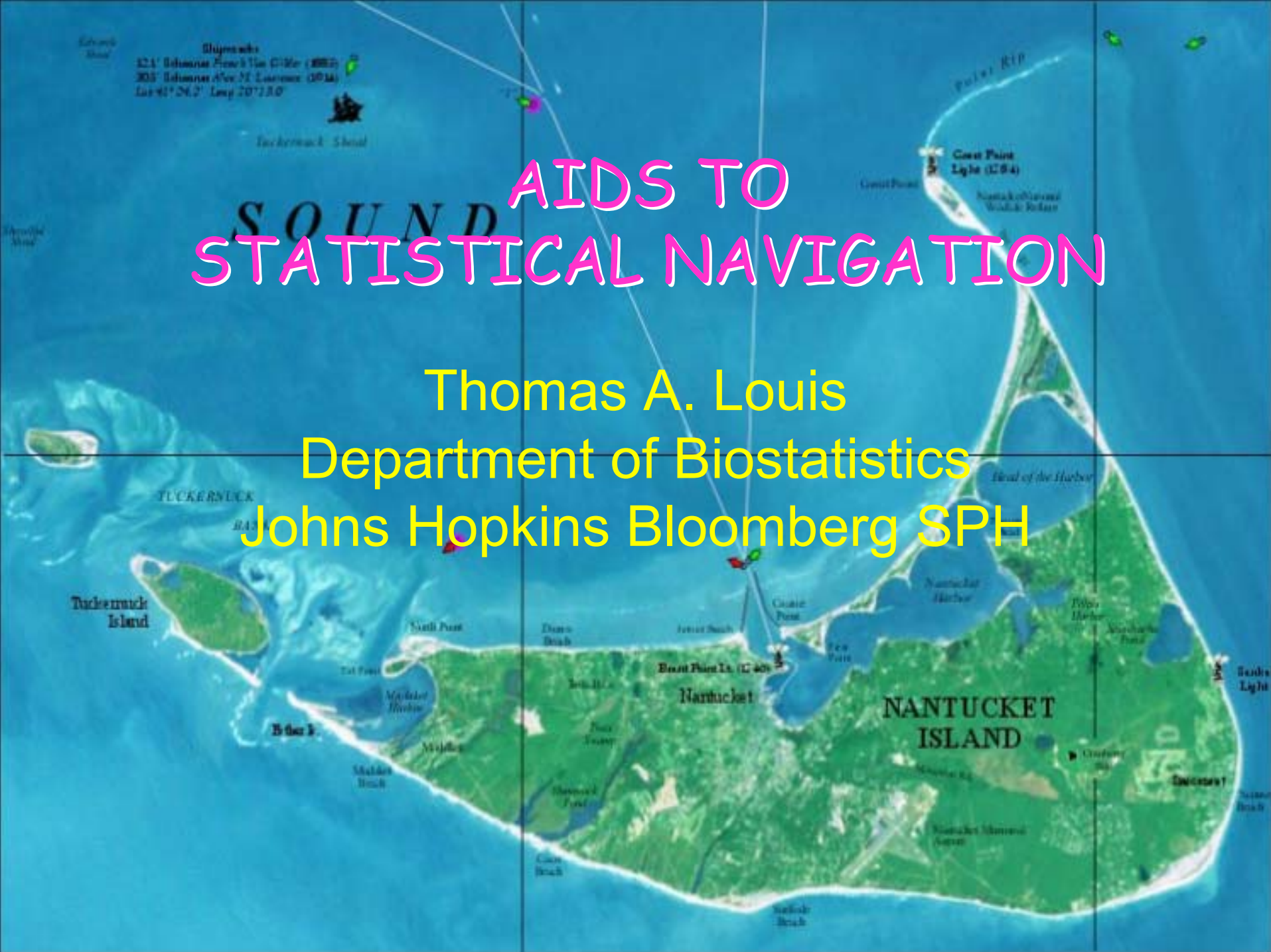


AIDS TO STATISTICAL NAVIGATION

Thomas A. Louis
Department of Biostatistics
Johns Hopkins Bloomberg SPH

Thanks to:

- Tim Gregoire
- My colleagues at Hopkins
- Brad Carlin, Patrick Heagerty, and Jiannong Liu

Support provided by:

- DHHS/NIH/NIDDK&NICHD,
- The United States Renal Data System
- The Carnegie Foundation

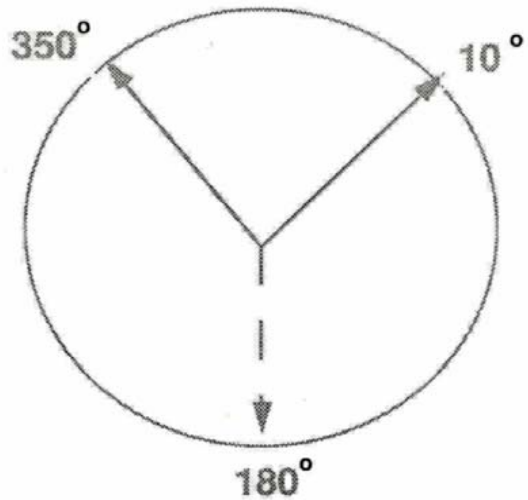
PREMISE

- Statistical practitioners employ a wide variety of designs and analyses
- By providing anchors and waypoints, an understanding of statistical philosophies (Frequentist, Likelihood, Bayes), stabilizes and smooths these variations
- However, valid and effective design, conduct and analysis require that this understanding be complemented and supplemented by a broad array of navigational aids
- I discuss the roles of a selected sample of philosophy-based and practice-based aids

TRADITIONAL VALUES

- Understand the application
- Relevant, accurate and timely data are fundamental

UNDERSTAND THE APPLICATION



Typical Direction

$$\frac{10^\circ + 350^\circ}{2} = 180^\circ \quad !$$

$$\frac{10^\circ + (-10^\circ)}{2} = 0$$

ACCURATE DATA

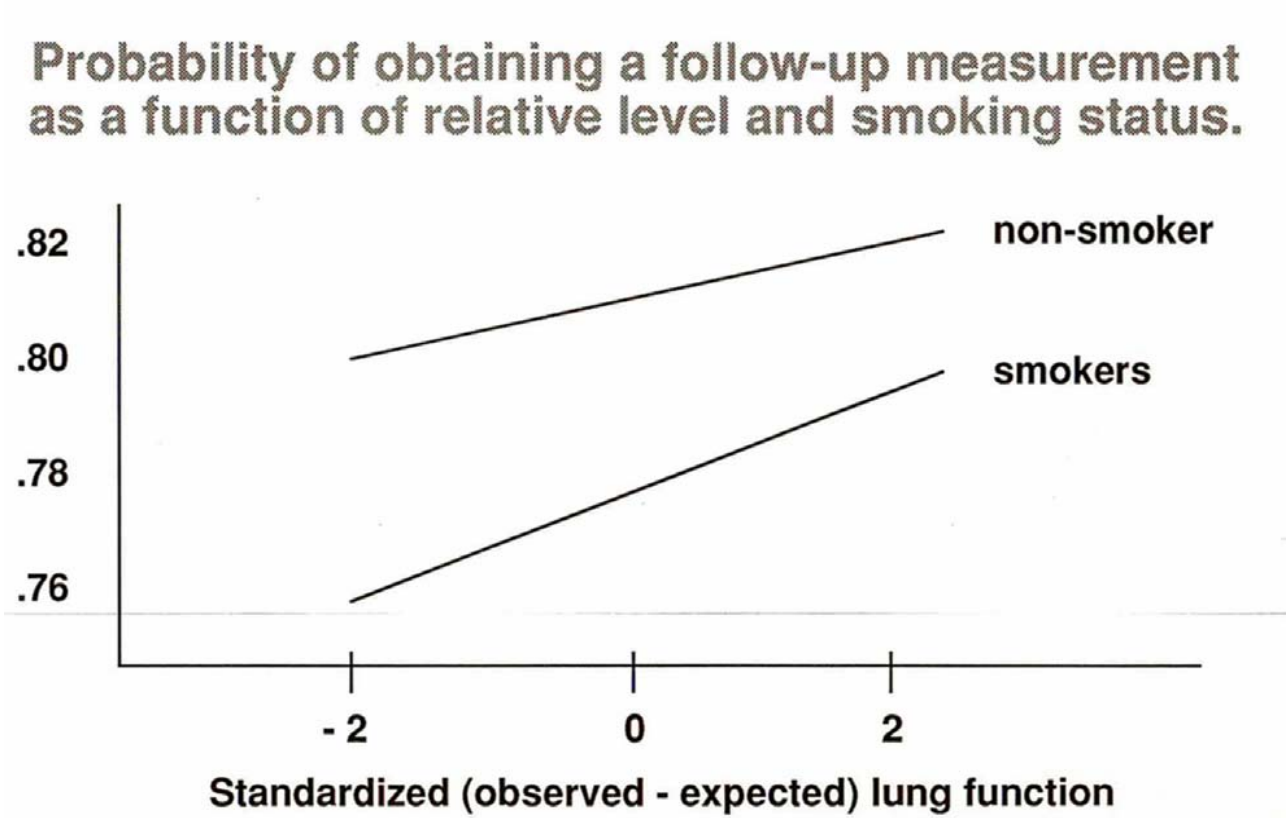
- RCT of a treatment designed to reduce side effects of chemotherapy
- Blinded interim results showed that it was almost impossible for the estimated treatment effect to reverse
- If the interim estimate favored the placebo group, it was time to stop the study; otherwise continue

THE ENVELOPE PLEASE

TRADITIONAL VALUES

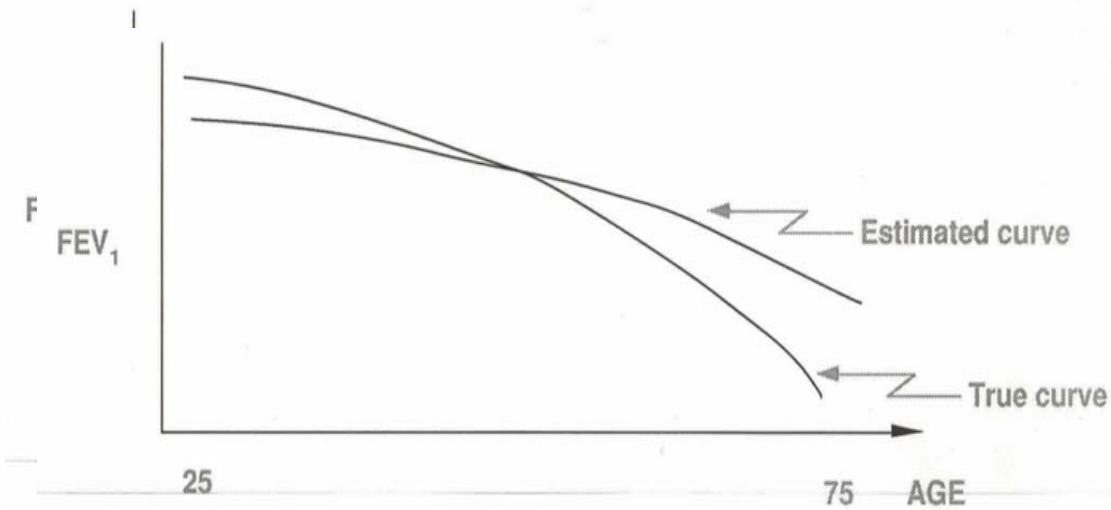
- Identify, incorporate and report stochastic and other uncertainties
- Identify goals and ensure that procedures serve them
- Identify and honor the sampling plan

SAMPLING PLAN
PROBABILITY OF FOLLOW-UP



SAMPLING PLAN

LUNG FUNCTION DECLINE IN ADULTS



Individuals are more likely to drop out if lung function decline is relatively steep or lung function relatively low.

“IGNORABLE” DROPOUTS?

- Ignorability is not an inherent property of the process
- It depends on the process and on the analytic model
- For example, MAR is ignorable, IF the analytic model is correct

FIXED EFFECTS OR RANDOM EFFECTS?

- Random effects inflate variance over fixed effects
- Addressing a broader inference space
  - "These studies" versus "Studies like these"
- Much of the FE/RE controversy concerns the inference space

ASPIRIN & CHD
(Canner 1983)

Study	100 OR	Z-Score	One-sided P-value	
CDPA	68	-1.90	.0290	
UK-1	72	-1.67	.0480	
UK-2	80	-1.53	.0630	
PARIS	80	-1.20	.1150	
GAMS	81	-.079	.2180	
Subtotal	76	-2.99	.0014	N = 3300
AMIS	113	1.27	.8980	N = 2300
TOTAL	90	-1.23	.1100	N = 5600

VARIANCE COMPONENTS
(log scale)

$$E[\log(OR_k)] = \mu,\ V[\log(OR_k)] = \tau^2$$

	without AMIS	with AMIS	Inflation Factor (%)
τ	.100	.180	80
$\hat{\mu}_{fixed}$	-.270	-.105	
$\hat{se}(\hat{\mu})$	.091	.086	-5
$\hat{\mu}_{rand}$	-.270	-.170	
$\hat{se}(\hat{\mu})$	.104	.138	33
Inflation Factor (%)	14	61	

"OPTIMALITY"

- Ensure that a procedure addresses your goal
- Don't confuse optimal with good
  - The "best of breed" may be a dog
- A procedure optimal for one goal may perform poorly for another

WEIGHTED MEANS
(CORRELATED REGRESSIONS)

- Unless their is a common expected value for all components of a weighted mean, the estimand will depend on the weights
- Unless a regression model is properly specified, the estimand depends on the working covariance
- “Optimal” weights may not be optimal

SUBOPTIMAL "OPTIMALITY"

$$E(\Sigma\, w_k Y_k) = \Sigma\, w_k\, \mu_k$$

- If $\mu_k \equiv \mu$ and $\Sigma\, w_k = 1$, then $E(\Sigma\, w_k Y_k) \equiv \mu$
- So, use variance minimizing weights

$$\boxed{w_k \propto \frac{1}{V(Y_k)}}$$

- If the $\mu_k \not\equiv \mu$, use "policy weights," e.g,

$$\boxed{w_k \propto (\text{population size})_k}$$

ROBUSTNESS & EFFICIENCY

- Balance robustness and efficiency
  - Buy reasonably priced insurance whenever possible
  - Be as flexible and robust as you can afford to be
- There is no free lunch in statistics
  - But, there are reduced-price lunches

BUY INSURANCE

- Small reductions in efficiency can produce substantial gains in robustness
  - Balanced designs
  - Efficiency of the Wilcoxon test/estimate
    - 96% for Gaussian; $>$ 86% for other distributions
- Three-point regression instead of two point
  - Allows diagnosing non-linearity
  - Costs about 12% in sample size

WORKING INDEPENDENCE?

- For longitudinal or spatial data, using a valid working covariance has the potential to increase efficiency
- And increase protection from data Missing At Random
- But, a non-independence working model amplifies the threat of bias due to an invalid fixed effects model
- Working independence protects against some of these misspecfications

INFORMATIVE SAMPLE SIZE
(via Patrick Heagerty)

- Y = birth weight
- X = time varying maternal PCB load
- Model the marginal mean: $E(Y \mid X, \text{live birth})$
- 300 potential mothers with up to 5 completed pregnancies
- Data generated by $Y = \beta_0 + \beta_1 X + \epsilon$
  - ϵ includes a random intercept, random slopes and auto-correlated errors
  - True parameters are $\beta_0 = \beta_1 = 1$

ESTIMATION METHODS

- Use the first birth only
- Use all births and GEE with working independence
- Use all births and a random intercept working covariance
  - GEE with working compound symmetry
  - Not the correct likelihood
- Robust SEs for all approaches

BABIES AT RANDOM

- Probability of a baby in the next time period does not depend on the past

	β_0	β_1
OLS, first-born only	$0.98_{(.37)}$	$0.99_{(.17)}$
GEE, working independence	$1.00_{(.04)}$	$1.00_{(.04)}$
GEE, random intercept	$1.00_{(.04)}$	$1.00_{(.04)}$

Simulation mean$_{(se)}$

- All approaches are unbiased
- Using only the first birth is inefficient

HISTORY-DEPENDENT BABIES

- The probability of having another birth depends on the previous birth outcomes

	β_0	β_1
OLS, first-born only	$1.05_{(.35)}$	$1.02_{(.18)}$
GEE, working independence	$1.01_{(.04)}$	$1.00_{(.04)}$
GEE, random intercept	$0.74_{(.06)}$	$\mathbf{0.91_{(.04)}}$

Simulation mean$_{(se)}$

- First-born and OLS are unbiased
- GEE (random intercept) has an incorrect estimand
  - And, in this example it does not reduce variance

THE ROLE OF FORMAL PHILOSOPHY

- The principal philosophies
  - Frequentist, Likelihood, Bayes, ...
- Provide a framework for practice
  - Are fundamental navigational aids
- Use them as guides; not as straightjackets
  - Kempthorne, “Don't bother me, I'm just looking at the data”

Savage, 1972

"As in other sciences, controversies over the foundations of statistics reflect themselves in everyday practice, but not nearly so catastrophically as one might imagine.

I believe that here, as elsewhere, catastrophe is avoided, primarily because in practical situations common sense generally saves all but the most pedantic of us from flagrant error."

PERSONAL OPINION

- Personal opinion pervades all of (statistical) science
- Identify, incorporate and communicate it
- Be formal whenever possible

MULTIPLICITY

- Personal opinion and goals
  (Prior distributions and loss functions)
  structure a multiplicity environment
- Do you want to control the Type I error for:
  - This assessment of this endpoint?
  - All assessments of this endpoint?
  - A group of your studies?
  - ⋮
  - Your research career?

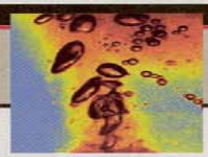
MULIPLICITY AND PERSONAL OPINION

- P = .0100 for the most significant of 10 independent comparisons
- What (adjusted) P-value should an investigator use?

Adjusted P-value	Rationale
.0100	I only planned to look at this one (it happened to be the most significant)
.0199	I planned to look only at 2 (one of which was the most siginficant)
⋮	
⋮	
.0956	I checked all 10

WOMEN'S HEALTH INITIATIVE

- Stopped in the Spring of 2002, due to an elevated risk of invasive breast cancer, CHD and Stroke
- The decision depended on not adjusting for multiplicity
- The British WISDOM Study continued
  - Based on a multiplicity adjustment and other considerations
- Janet Wittes, "...in the end, it's often a difficult decision"



PAGE 496
Bubble fusion deemed unlikely

497
Forecasting El Niño

EPIDEMIOLOGY

Despite Safety Concerns, U.K. Hormone Study to Proceed

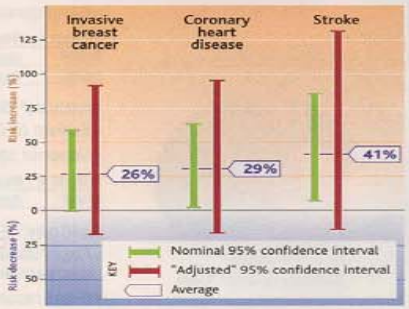
When a huge U.S. study of hormone drugs came to an abrupt standstill 2 weeks ago, women's health experts cast a curious glance across the Atlantic. Would leaders of an even bigger trial funded by the U.K. government and the Medical Research Council be shaken by the damning evidence of risk from the Women's Health Initiative (WHI)? Would they halt their study and advise participants to stop taking their pills? Not at all, is the surprising answer.

Meeting last week, an independent safety panel for the $32 million Women's International Study of long Duration Oestrogen after Menopause (WISDOM) unanimously concluded that WHI's evidence that hormone replacement therapy raises the risk of heart disease is not convincing. The trial's steering committee, equally skeptical, decided to forge ahead with WISDOM. The U.S. researchers "have not determined the size of the risks reliably," says the chair of the steering committee, Oxford epidemiologist Rory Collins. This split, according to many observers, reflects a difference between cultures as much as a disagreement over the science.

The British researchers argue that WHI's ambiguous results make WISDOM even more important. Proponents of hormone therapy also take heart in the prospect that WISDOM's fresh data might offset the recent findings. But the decision is drawing criticism from some U.S. researchers, who say the study puts women at unnecessary risk. "It may be good for science," says epidemiologist Curt Furberg of Wake Forest University in Winston-Salem, North Carolina, "but patients will pay the price." WHI director Jacques Rossouw calls the British misgivings about his study "bogus."

Both WHI and WISDOM were designed to test a combination of estrogen and progestin. These hormone substitutes are taken by millions of women to counter short-term menopausal symptoms such as hot flashes and insomnia, or to prevent age-related diseases over the long term, such as osteoporosis and heart attacks (*Science*, 19 July, p. 325). More than 16,000 U.S. women participated in WHI; WISDOM has enrolled 5000 British women so far but aims to recruit 17,000 more in the United Kingdom, Australia, and New Zealand.

The U.S. trial ended when an independent data and safety monitoring board (DSMB) concluded 31 May that the increased risk of breast cancer from hormones had crossed a previously set line. Typically, DSMBs study the accumulating data several times a year and recommend quick action if early results show that a drug is clearly effective—making it unethical to continue participants on placebo—or clearly harmful. Statistics can help clarify the choices, but "in the end, it's often a difficult decision," says statistician Janet Wittes, president of the Statistics Collaborative in Washington, D.C., and chair of the panel that has monitored WHI since 1997. In the case of WHI, the data failed to convince WISDOM's counterpart review panel, chaired by statistician Richard Gray of the University of Birmingham, U.K.



Uncertainty. The range of likely risk from hormone therapy is wider when the data are adjusted for multiple sampling.

The two teams look at risk differently. In their paper in the 17 July issue of the *Journal of the American Medical Association* (*JAMA*), the U.S. researchers present the average risks for each adverse outcome along with a "confidence interval"—a range of values between which the true value lies with 95% certainty. But the approach has a problem, says Gray: Analyzing data multiple times and testing for several outcomes increases the chances of getting a spurious result.

The WHI team also presented "adjusted confidence intervals," statistically corrected for multiple sampling, which show a different picture. For each of the three main adverse outcomes, the adjusted range includes the possibility of no increased risk at all or even a decreased risk (see graph). That leaves room for doubt about hormones' effect on heart disease, says Gray—especially because the increased risk conflicts with previous data. (His panel does not question the findings on stroke and breast cancer.)

To Rossouw, this is a quibble. He says the adjusted intervals were put in the *JAMA* paper mainly as a service to "aficionados," adding that most trial reports never even mention them. He thinks the WISDOM researchers are looking for "an excuse to dismiss the results. ... They have their own trial to protect."

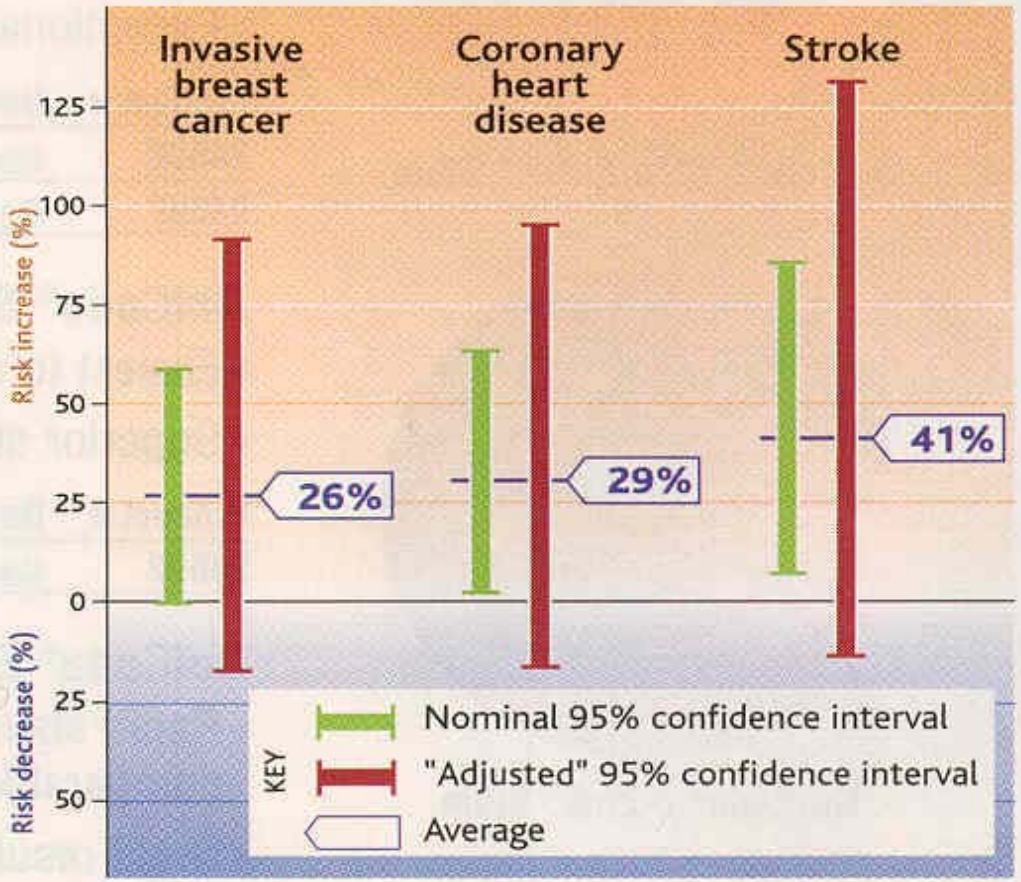
The rift highlights how ethical frameworks differ on opposite sides of the Atlantic, others say. Whereas U.S. DSMBs tend to emphasize individual patients' safety, "the Brits are much more comfortable proceeding with a trial in order to get statistically significant results," says Wittes. Collins points out, however, that stopping a trial early is not always the best way to protect patients. He recalls that several U.S. trials of AZT, the first AIDS drug, were cut short because it seemed to prolong life. But despite pressure to halt, a French-British collaborative study, called Concorde, continued and eventually showed that AZT alone had no overall effect on mortality.

What remains unclear in this technical debate is how women in the U.K., Australia, and New Zealand will respond. Will they want to enroll in WISDOM now that American participants have been advised to stop taking the pills? All current and future WISDOM participants will be told about the U.S. results, says Collins, and some might opt out. But it's possible that the dropouts will be offset by newly motivated recruits. In the past, some women were so convinced about the benefits of therapy that they refused to participate, he says. Some might now enroll in WISDOM to support a second look at that question.

–MARTIN ENSERINK

SOURCE: *JAMA*

president of the Statistics Collaborative in



Uncertainty. The range of likely risk from hormone therapy is wider when the data are adjusted for multiple sampling.

EVIDENCE & DECISIONS

- Goodman and Royall, 1988
  "Evidence is a property of data that makes us alter our beliefs about how the world around us is working"
- Ideal analysis sequence
  - Prior knowledge/beliefs
  - Evidence (need to decide what is evidence)
  - Inferences/decisions (via loss functions)
- Likelihood and Bayesian structuring
  - To be likelihood based, you need a likelihood

SIGNIFICANCE TESTING

(Richard Royall)

- $X \sim$ Binomial(30, θ)
- $H_0 : \theta = \frac{1}{4} \quad H_1 : \theta = \frac{3}{4}$
- The Likelihood Ratio (H_1 vs H_0)

$$\boxed{LR(x) = LR_{30}\left(\tfrac{3}{4}, \tfrac{1}{4}, x\right) = 9^{x-15}}$$

$$LR(x) = LR_{30}\left(\tfrac{3}{4}, \tfrac{1}{4}, x\right) = 9^{x-15}$$

- The 5% level test of $\theta = \frac{3}{4}$ versus $\theta = \frac{1}{4}$
  Rejects for $x \geq 12$
- But, $LR(12) = 1/729$
- So, the test rejects $\theta = \frac{1}{4}$ at an x-value for which it is favored $(729 : 1)$ over $\theta = \frac{3}{4}$
- This decision rule can only be justified by
  - A prior belief that $pr(\theta = \frac{3}{4})$ is large
  - A large $(\text{loss}_1)/(\text{loss}_0)$

CONFIDENCE INTERVAL AFTER OBSERVING 0 EVENTS

- Observe $X = 0$ in n trials
- The frequentist, one-sided $100(1-\alpha)\%$ CI is

$$\left[0, \frac{X_{future}}{n}\right) = \left[0, \frac{-log(\alpha)}{n}\right)$$

- For $\alpha = 0.05$ is $\left[0, \frac{3}{n}\right)$
- For many people, "$X_{future} = 3$" is the smallest number of pseudo events that seem "unlikely"

CALIBRATING LRs AND BAYES FACTORS

$$\boxed{e^x = LR_n\left(\theta = \tfrac{x}{n},\ \theta = 0\right)}$$

- For $x = 2, LR_n = 7.4$
  - A moderate evidence LR and Bayes factor
- For $x = 3, LR_n = 20.0$
  - A strong evidence LR and Bayes factor

THE ROLE OF RANDOMIZATION

- Produces statistical independence between attributes and treatment assignment
  - The propensity score is known
  - Enabling causal inferences
- Provides a basis of inference
  - The SE takes care of the possibility of imbalance
  - Randomization tests
- Enhances credibility

DESIGN- BASED INFERENCES?

- Randomization tests are valid
- But, conclusions depend on the randomization plan
- Do you want to come to different conclusions if one trial used 50/50 and another used 70/30 probabilities, but had identical results?

IS UNBIASEDNESS A WORTHY GOAL?

- "Bias" is an unfortunate statistical term because it has a strong negative connotation
- In statistics, bias is bad only if it connotes "deceptive" or in other ways "unscientific/unethical"
- It does produce "best unbiased" frequentist solutions
  - But, the "best of breed" can still be a dog

GOAL SHOULD BE "GOOD ESTIMATES"

- MSE $=$ Variance $+ \left(\text{Bias}\right)^2$
- A little bias can pay big variance reduction dividends
  - C_p , Adjusted R^2, PRESS residuals, ...
  - Soft/hard thresholding
  - Shrinkage estimates

SHRINKAGE ESTIMATES

- Strike an excellent Variance/Bias trade-off
  - Stabilize high variance estimates more than low variance ones
  - Partial inclusion of regressors
  - Partial inclusion of residuals
    - Going part way from a fixed effects model $(\mathbf{X}\boldsymbol{\beta})$ to the data
  - Partial de-attenuation

BAYESIAN INFERENCE

- Specify the structure
- Compute the joint posterior distribution for ALL unknowns
  - The Rules of Probability dictate how
  - Doing it can be very difficult [:–(& (–:]
- Compute the marginal posterior distribution for features of interest
  - Parameters, missing data, future data, ...
- Use the posterior to make inferences
  - Squared-error loss $(a - \theta)^2$
    $\Rightarrow \hat{\theta} = E(\theta \mid Y)$ (the posterior mean)

SCOTTISH LIP CANCER DATA

- Data on the 56 counties, pooled over 1975-1980
  - The male population and years at risk
  - Observed and expected lip cancer cases
    - ◇ internal standardization
  - A covariate measuring the *fraction* of the population engaged in agriculture, fishing or forestry (AFF)

A SPATIAL HIERARCHICAL MODEL

$$\begin{aligned}
\boldsymbol{\eta} &\sim h(\boldsymbol{\eta}) \\
\boldsymbol{\psi} &\sim g(\boldsymbol{\psi} \mid \mathbf{X},\ \boldsymbol{\eta}) \\
y_k \mid \psi_k &\sim Poi(m_k \psi_k) \\
\log(\psi_k) &= \mathbf{X}_k \boldsymbol{\alpha} + \theta_k
\end{aligned}$$

- The $\boldsymbol{\alpha}$ are *fixed effects*
- The θ_k are, possibly correlated, region effects (the Census Bureau calls them "model lack of fit")
- Evaluate adjusted relative risk: $\rho_k = e^{\theta_k}$

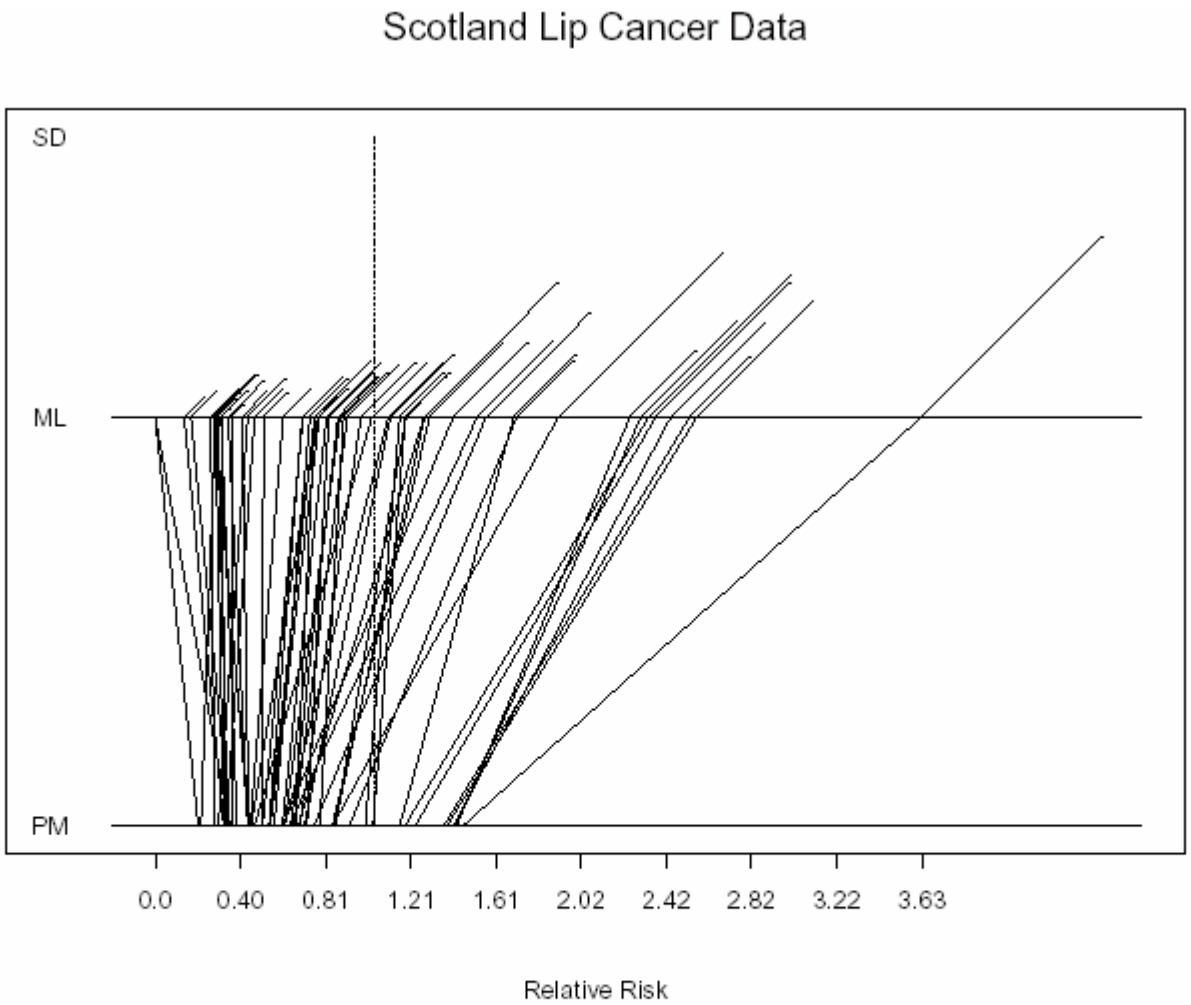


Figure 1: Maximum likelihood estimates (ML) of relative risk, their standard errors (SD) and posterior mean (PM) relative risks computed from the exchangeable model

RANKING

Birmingham Post January 23, 2003, Thursday

LEAGUE TABLES: AN ALL-BOYS GRAMMAR IS THE 'BEST IN ENGLAND'

The Independent January 23, 2003, Thursday

EDUCATION: SCHOOLS FACE CLOSURE OVER POOR RESULTS;
LEAGUE TABLES

OPTIMAL RANKING
(when variances are not equal)

- MLE-based ranks
  - o Unfairly declare relatively high variance estimates as extreme
- Hypothesis test-based ranks
  - o Unfairly declare relatively low-variance estimates as extreme
- A loss function is needed to structure the analysis

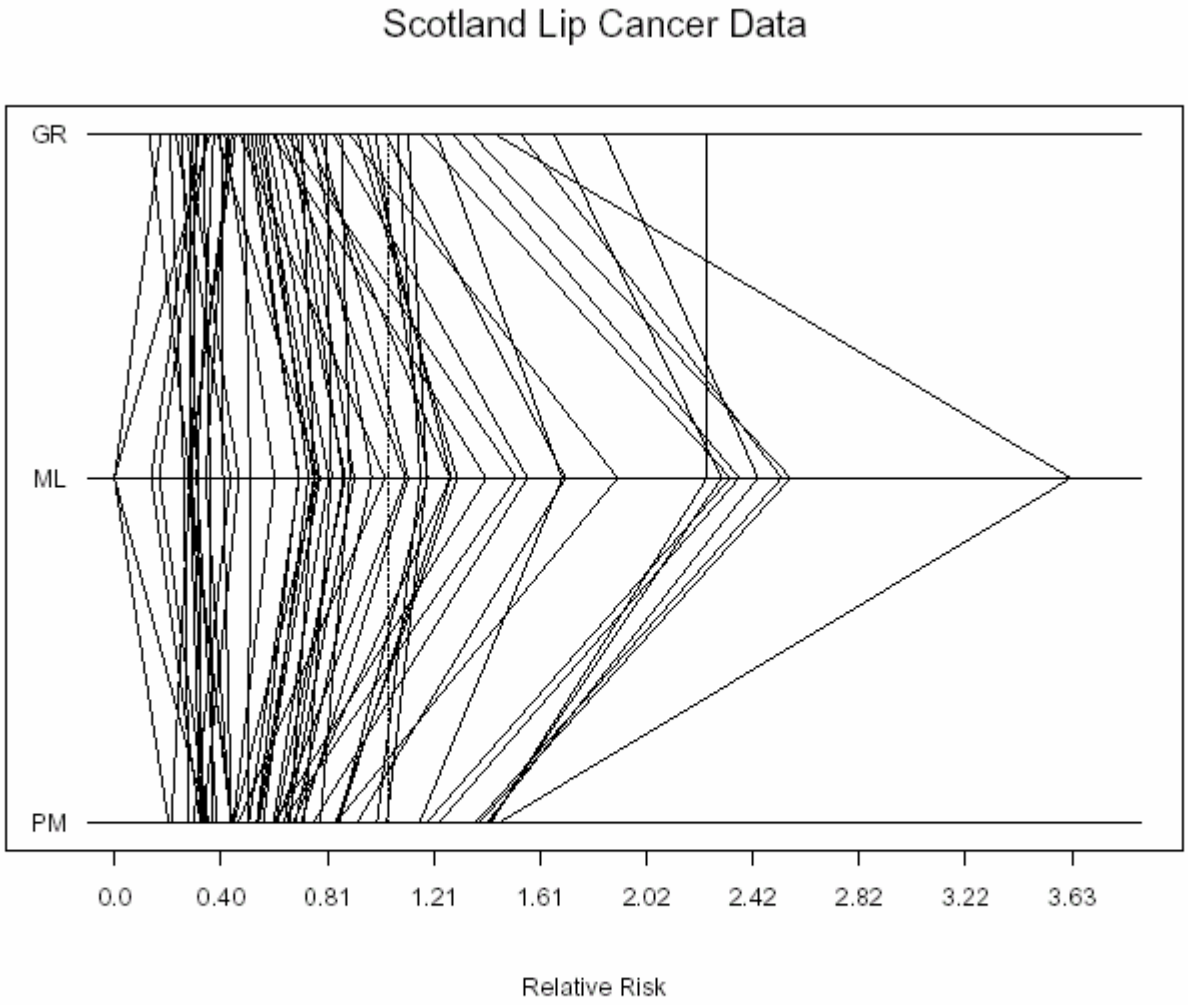


Figure 6: GR, ML, and PM estimates of relative risk for the Exchangeable model.

RANKING SMRs
(USRDS Jiannong Liu)

- 3459 dialysis providers in 1998
- Ranging from small (1 patient per year) to large (355 patients per year)
- Standardized mortality ratio

$$\text{SMR}_k = \frac{\text{observed}_k}{\text{expected}_k}$$

  - Expecteds are from a Poisson mixed effects, case mix adjustment model

- Rank the SMRs

ATTRIBUTING EVENTS

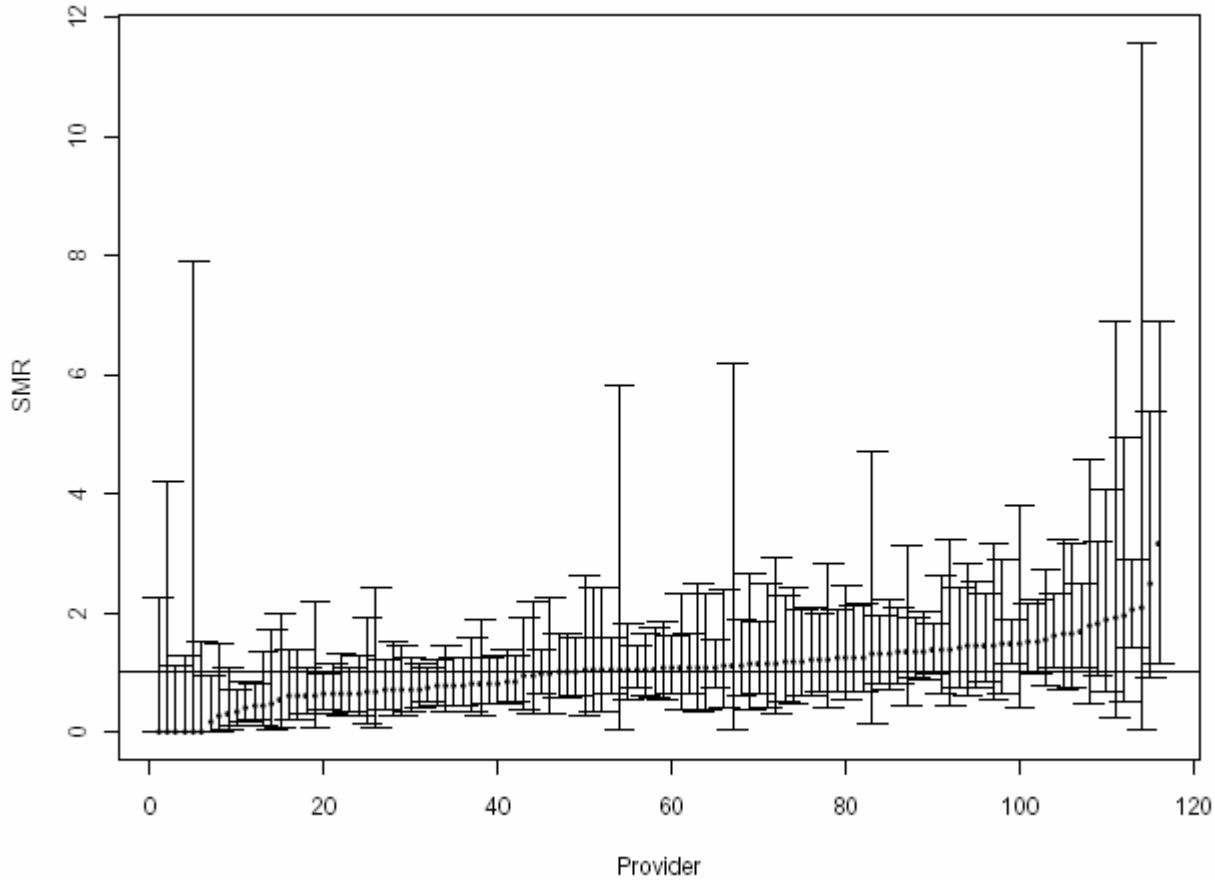
- Only while on this treatment
- Only while in this dialysis center

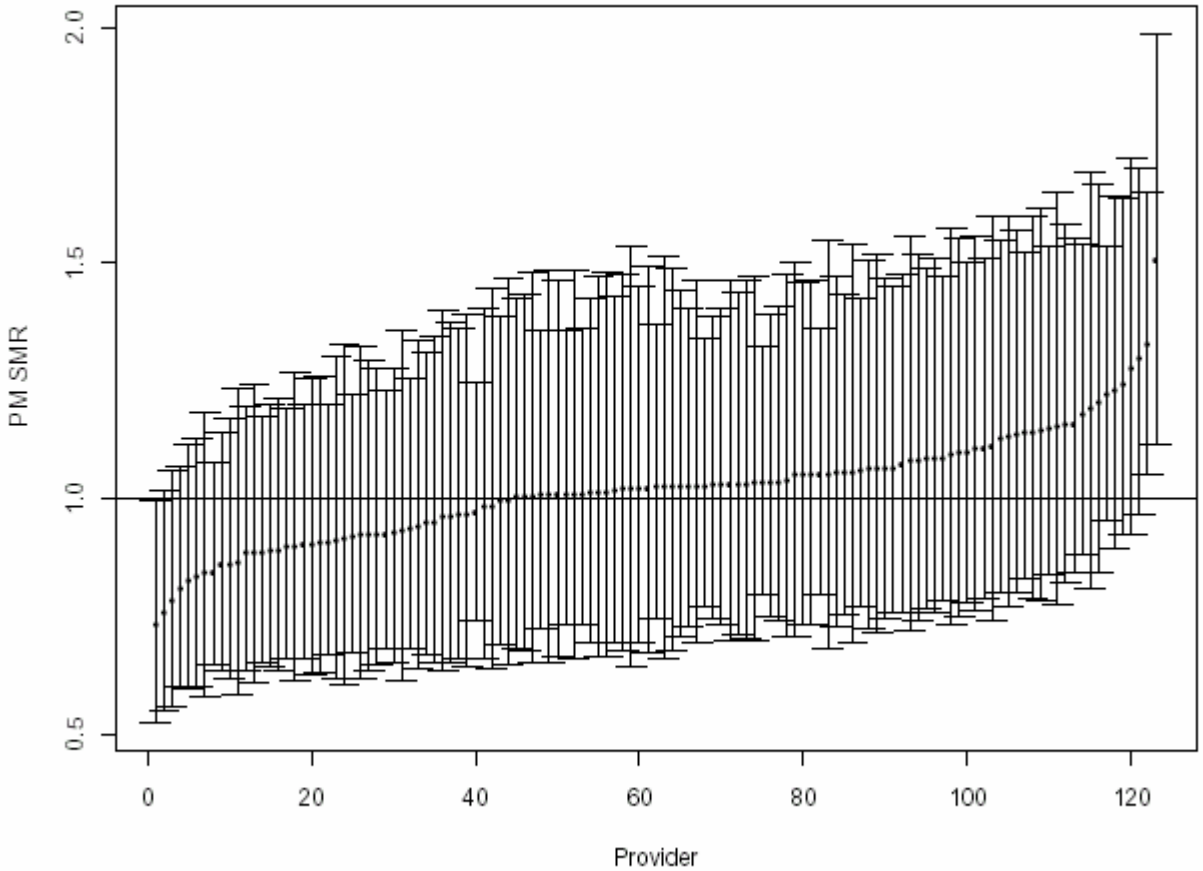
or

- For ## months after a treatment switch or a change in center

POSTERIOR MEAN BASED RANKS

- Ranks based on the PM (of the θs) aren't optimal
- But, because the PMs (almost) balance variability, they produce good ranks


SMR
0
2
4
6
8
10
12
0
20
40
60
80
100
120
Provider


PM SMR
2.0
1.5
1.0
0.5
0
20
40
60
80
100
120
Provider

BAYESIAN DESIGN

- Everyone is a Bayesian in the design phase
- Bayesian design for frequentist analysis
- Bayesian design for Bayesian analysis

BAYESIAN DESIGN FOR FREQUENTIST ANALYSIS

- θ is the mean of a Gaussian distribution
- $H_0 : \theta = 0$ versus $H_a : \theta > 0$
- Sample size to achieve a desired power (β) at θ_* is:

$$n(\theta_*, \sigma) = \left(\frac{\sigma}{\theta_*}\right)^2 (Z_\alpha - Z_\beta)^2$$

- Can do "what if" analyses wrt σ
- Or, design based on a distribution for it

THE BAYESIAN WAY

- Let $\sigma \sim g$; "Power" is a random variable
- Pick $n_g = n_g(\theta_*)$ to produce a required expected power:

$$E[\text{power}(n_g, \sigma)] = \int \Phi\left(\frac{\theta_* \sqrt{n_g}}{\sigma} + Z_\alpha\right) g(\sigma) d\sigma = \beta$$

- Or, pick n_g to make the probability of high power high:

$$pr[\text{power}(n_g, \sigma) > \beta] \geq \gamma$$

- Generally, $n_g(\theta_*) > n(\theta_*, \sigma_{typical})$

PROCEDURE GENERATION AND EVALUATION

- Be creative and eclectic in developing and implementing procedures
  - o Model based & Philosophy based
  - o Algorithmic
- Be scientific in evaluating them, guided by statistical philosophies
- Evaluate for “scenarios” of interest
  - o Generating processes
  - o Evaluation Criteria

ALGORITHMS & MODELS

- Early in their career, most procedures are “algorithms” and are not model based
- The “good ones” will have a model basis
  - o Wilcoxon,Mantel Haenszel, Lasso
  - ? Slow learing, SVM, ...

MOST ARE HYBRIDS

- Most Analytic strategies are hybrids
  - Composed of model-based components
  - With substantial algorithmic aspects
- Gene expression: microarrays and SAGE
  - Standardize intensities, correct for background, summarize perfect matches and mismatches, estimate the error rate, de-shadow, de-noise

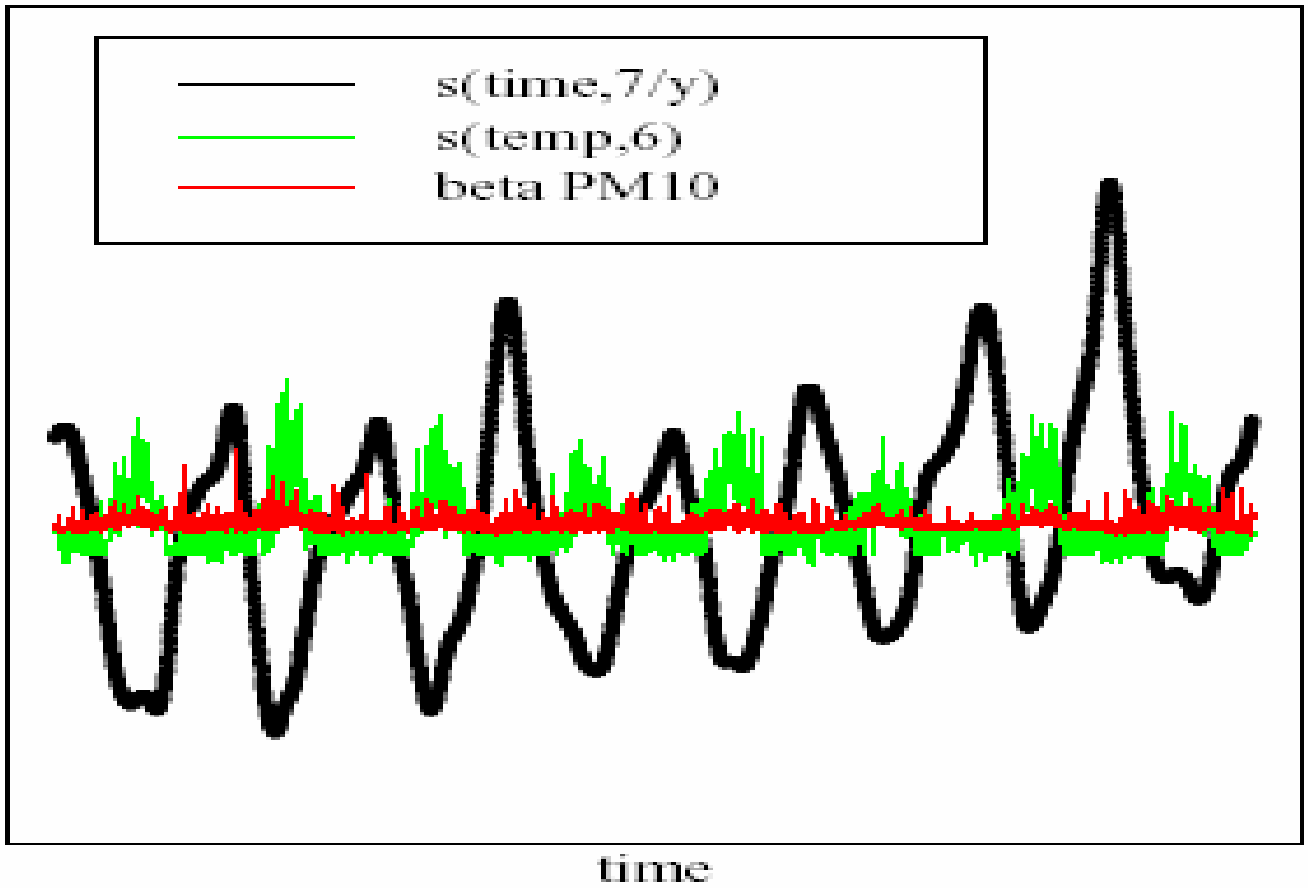
NMMAPS

National Mortality and Morbidity Air Pollution Study

(via Francesca Dominici)

- Estimate the percentage increase in mortality associated with a 10 unit increase in particulate matter
- Adjust by confounding factors such as weather and seasonality
- Challenging because the air pollution effect (the signal) is embedded in substantial confounding (the noise)
- So, must employ complex statistical adjustment models that extract the signal

Air pollution signal small relative to confounders



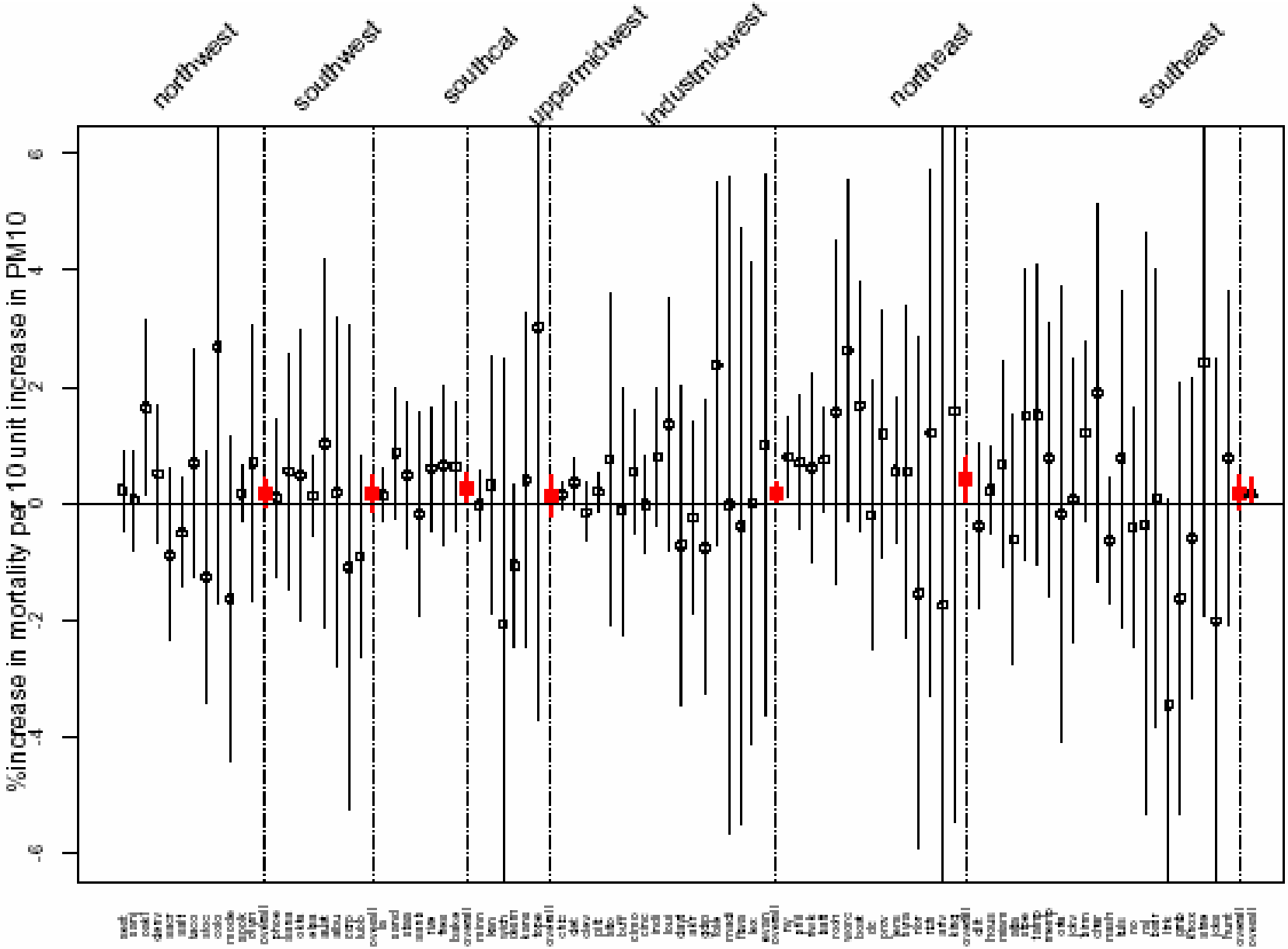
Estimates of model predictors in the GAM model Pittsburgh (1987-1994)

CONFOUNDNG ADJUSTMENT

$$
\begin{aligned}
Y_t &= \text{deaths at time } t \\
X_t &= \text{lagged air pollution} \\
Y_t &= \beta X_t + \text{splinesmooth}(t, df) + \epsilon_t
\end{aligned}
$$

- Strategy:
  - Remove long term associations between Y_t and calendar time
  - Study short term associations with X_t
- β is the adjusted air pollution effect
  - Its meaning does not depend on df

City-specific and regional estimates



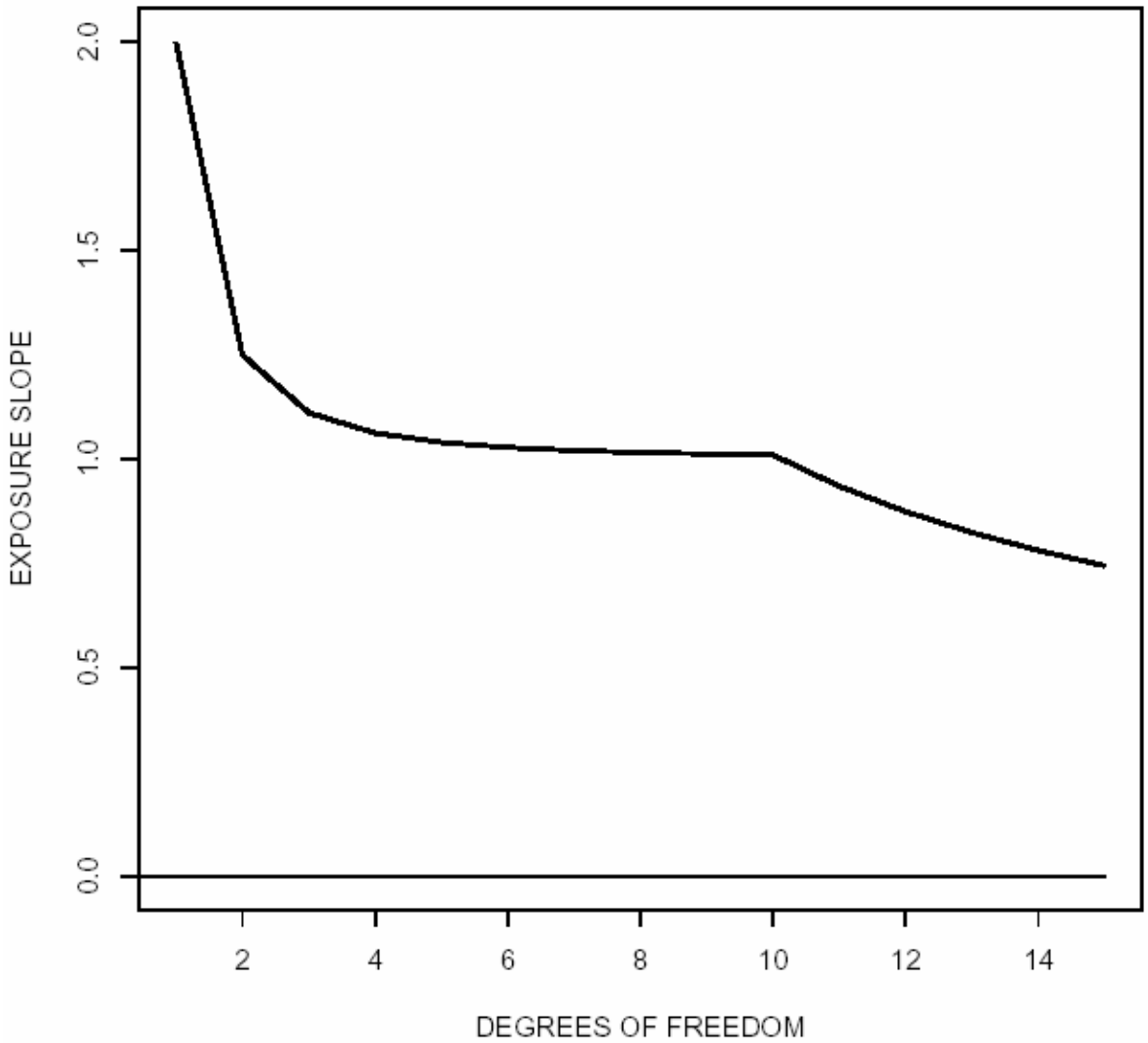
INCORPORATING AND COMMUNICATING UNCERTAINTY

- Accompany each statistical summary with informative measures of uncertainty
- Measure and incorporate dominant uncertainties
  - o Working model-based
  - o Model uncertainty
- Communicate dominant uncertainties and their consequences

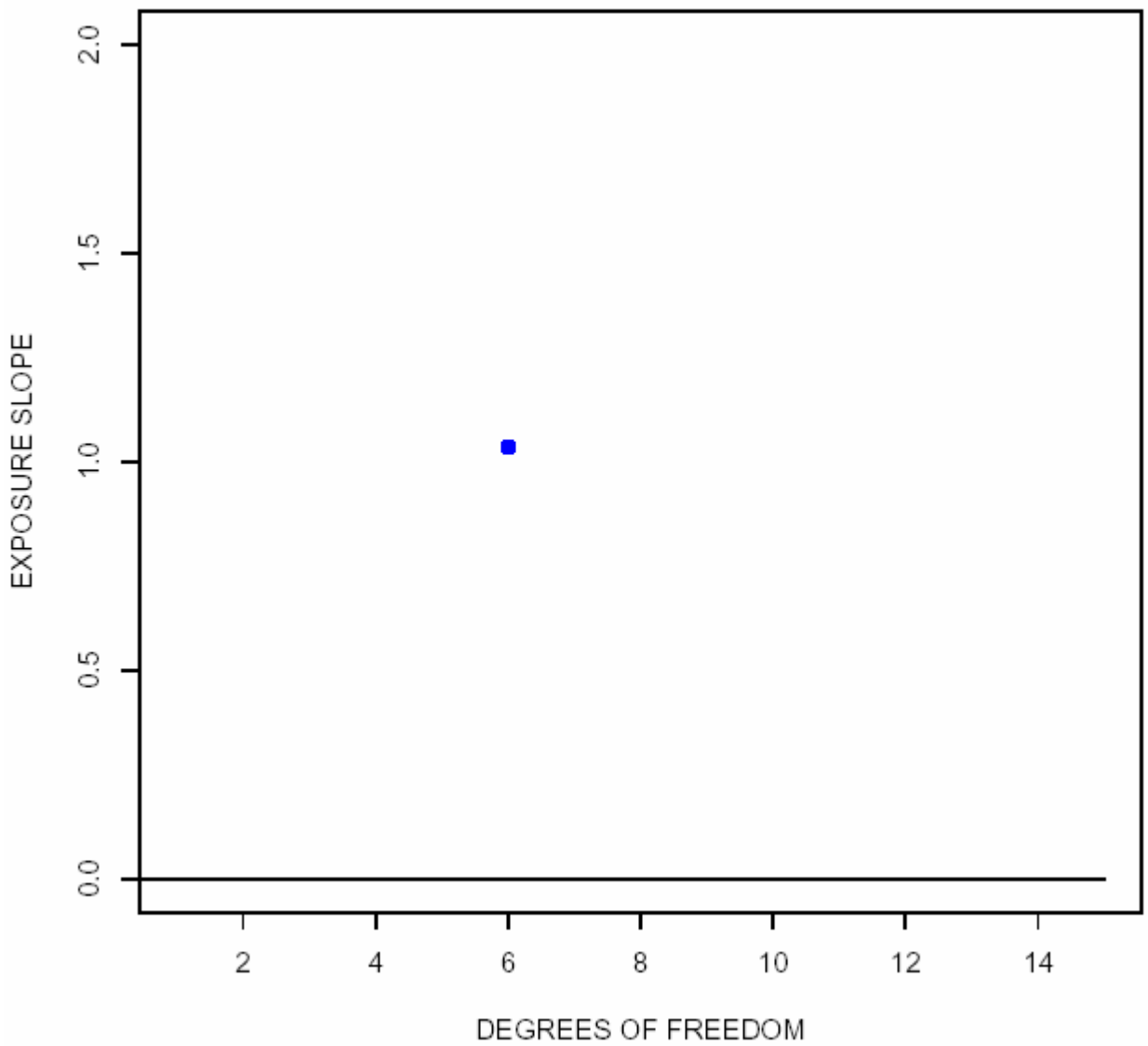
SELECTING DEGREES OF FREEDOM

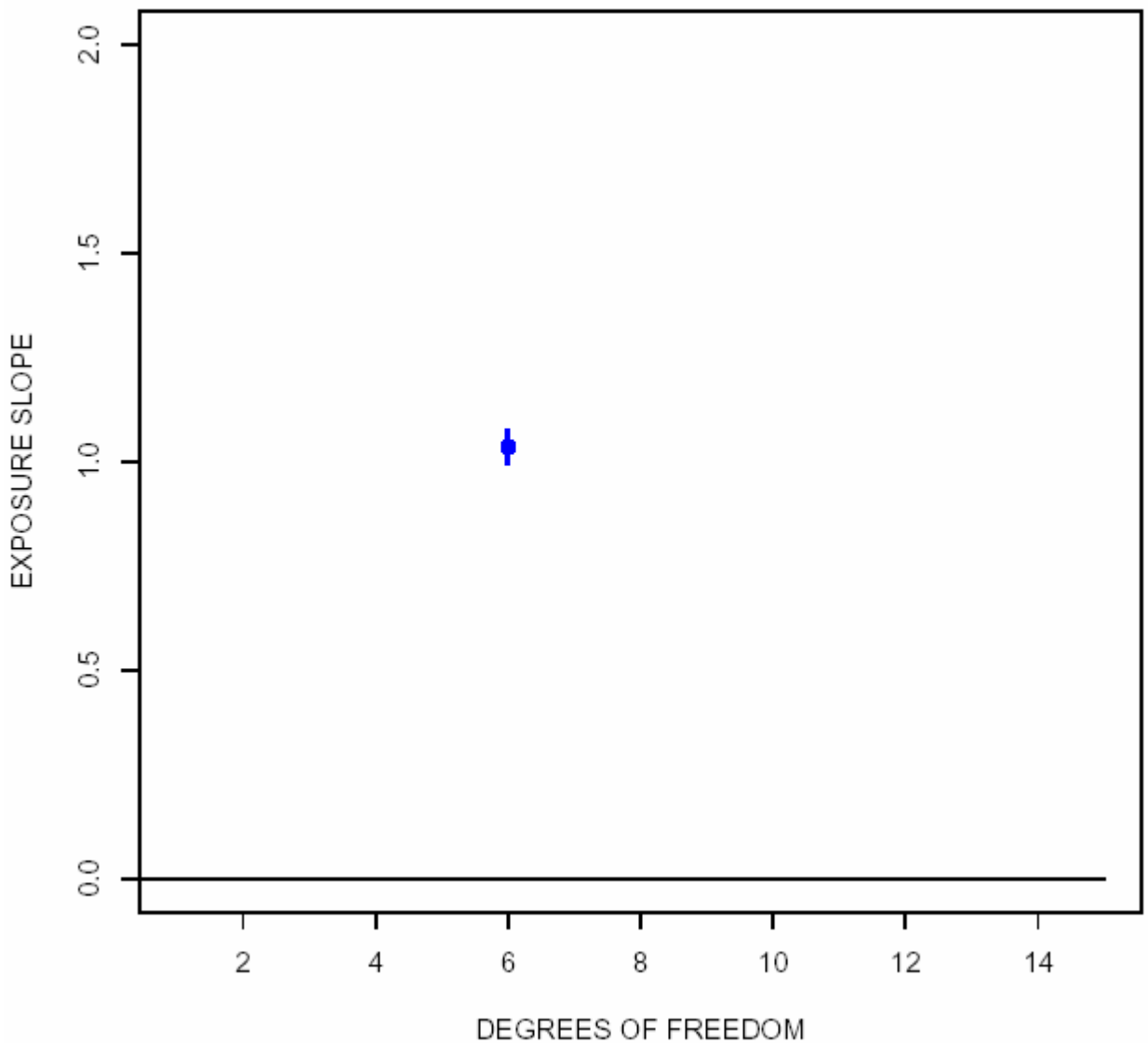
- Frequentist
  - AIC, BIC, sensitivity analysis
- Bayesian Model Averaging
  - Put a prior on df
  - Compute the full posterior for df and other parameters
  - Model average and focus on β
  - Report summary and uncertainty

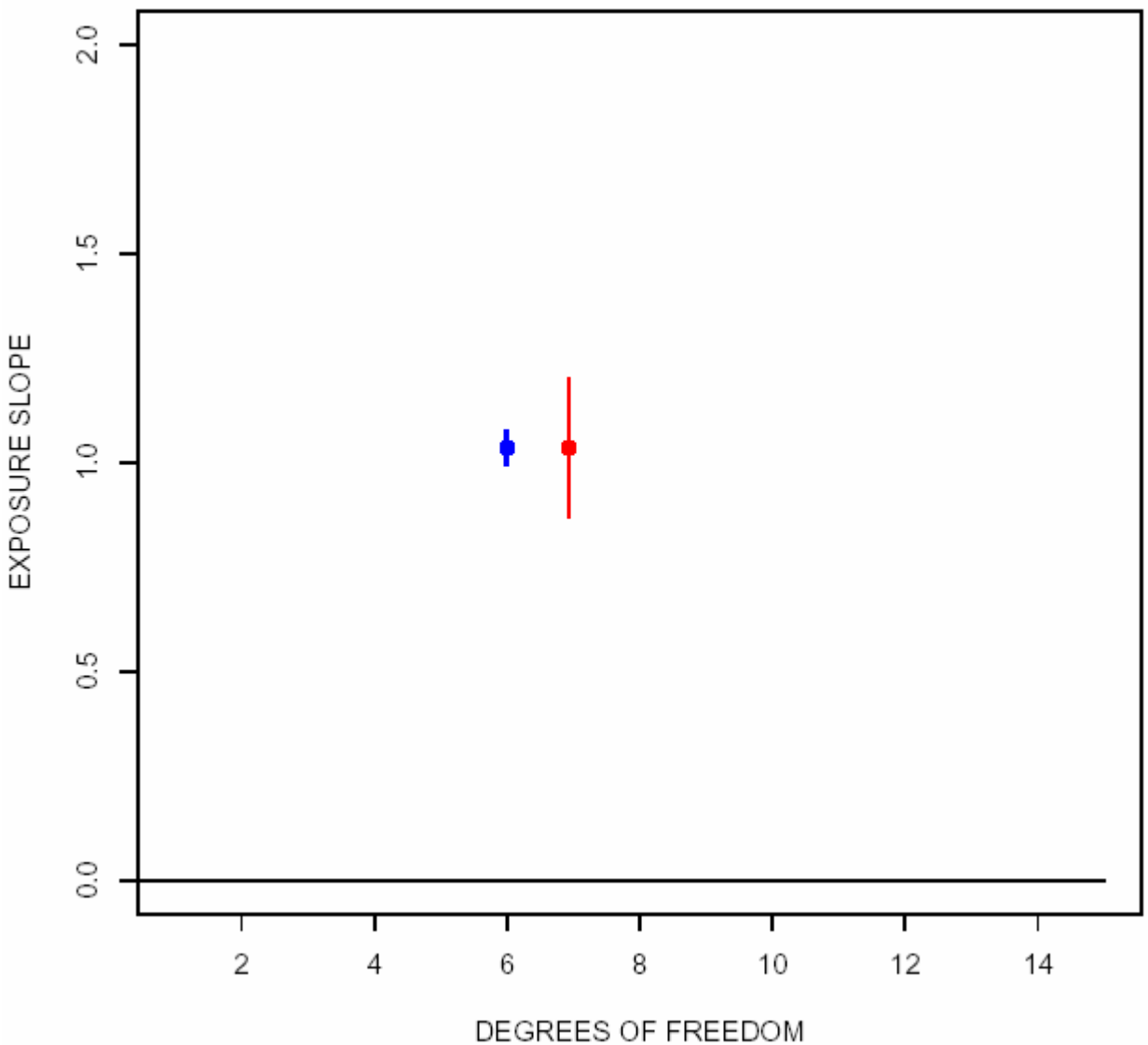
SLOPE ESTIMATE vs DF

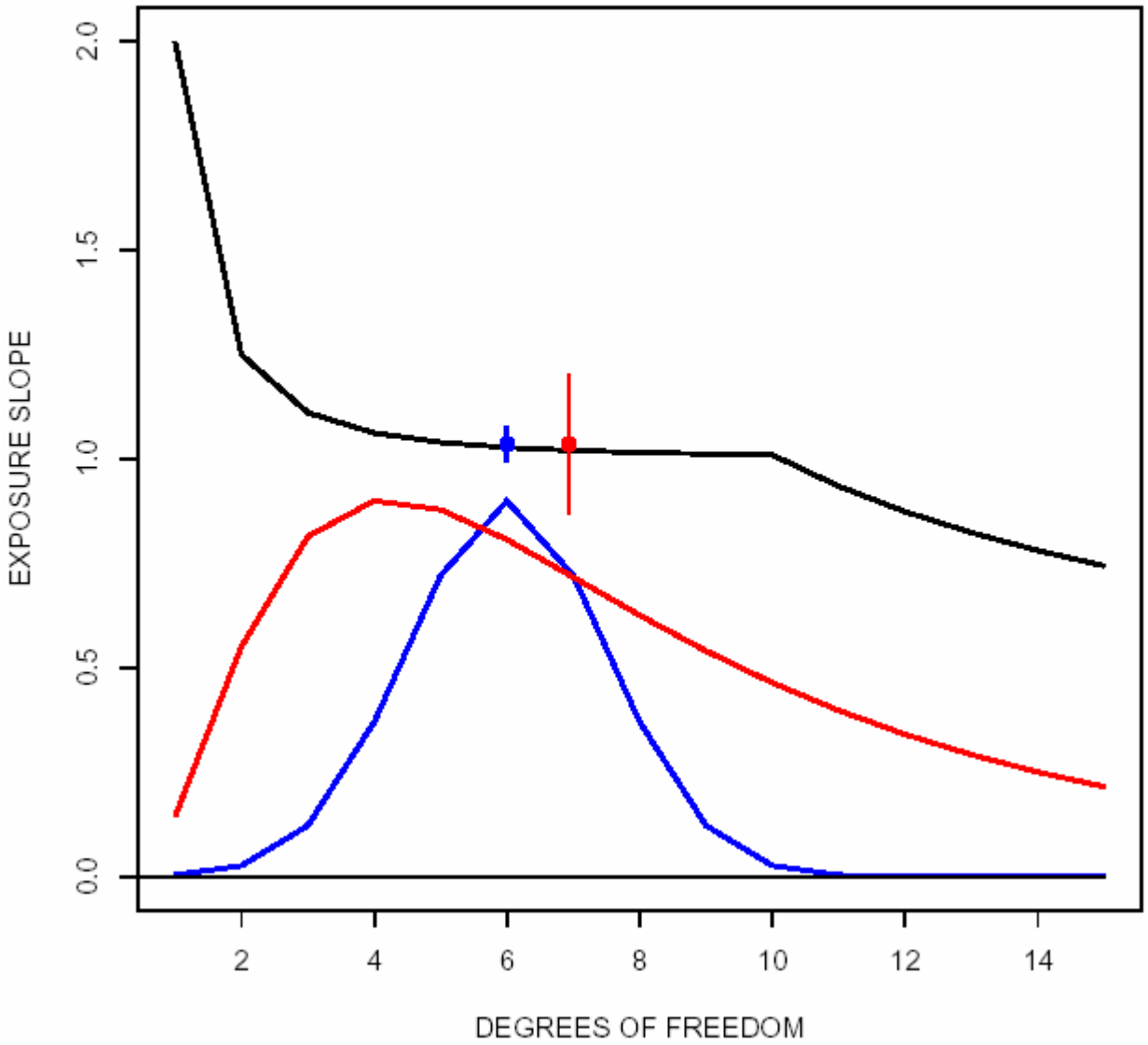


BAYESIAN MODEL AVERAGING


EXPOSURE SLOPE
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
12
14
DEGREES OF FREEDOM


EXPOSURE SLOPE
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
12
14
DEGREES OF FREEDOM


EXPOSURE SLOPE
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
12
14
DEGREES OF FREEDOM


EXPOSURE SLOPE
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
12
14
DEGREES OF FREEDOM

PARTICIPATING IN HIGH-CONSEQUENCE STUDIES

- Commit sufficient time and other resources
- Responsibly communicate findings and uncertainties in the simplest terms possible
- Identify and address sociological, psychological, economic and political influences
- Be prepared for contention

With the goal of high quality science and policy

THE GAM SLAM

- NMMAP estimates were produced using Generalized Additive models
- Default convergence criteria were too lax and estimates based on more stringent convergence criteria were smaller
- Creating quite a stir at the EPA and elsewhere

The New York Times, June 6 2002

Data Revised on Soot in Air and Deaths

Scientists Lower Their Estimate of Risk From Bad-Air Days

Revisiting their own data with new methods, scientists who conducted influential studies that linked sooty air pollution with higher death rates have lowered their estimate of the risk posed by bad-air days.

New research may delay a review due next year on small-particle pollution.

The researchers, at the Johns Hopkins University, have been distributing their new analysis to scientists and government officials by fax and e-mail. Yesterday, they set up a Web site, biosun01.biostat.jhsph.edu/~fdominic/research.html, that details their new findings.

Dr. Zeger and Mr. Greenbaum stressed that the new findings did not overturn the basic link between soot and illness. They also pointed to the recent publication of other studies on the long-term effects of soot that do not use the same analytical tools.

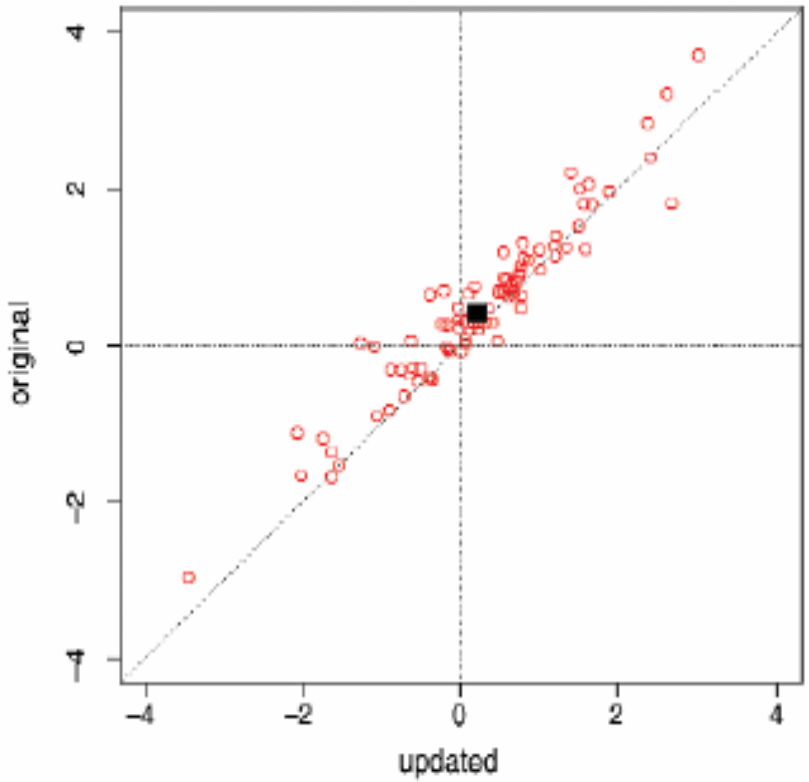
Still, industry officials said the new findings called into question the validity of some research underlying the new federal standards.

Science June 10 2002

AIR POLLUTION RISKS

Software Glitch Threw Off Mortality Estimates

The authors of a landmark air pollution study have found a problem with their software application that means they overestimated the risks of fine particles, or soot. The overall conclusions of the group at Johns Hopkins University in Baltimore linking soot and death haven't changed, but the discovery is providing fresh ammunition to industry groups that have criticized the science behind federal air pollution rules issued 5 years ago. The Environmental Protection Agency (EPA) says it will examine whether the rules need to be modified to reflect the new results.



Recalculating the risk. In this reanalysis of air pollution data, the vertical distance of dots from the diagonal line shows how much the estimated excess death rate was off for each of 90 cities. Black square represents updated (0.27% per 10 μg/m³ of PM_{10}) and original (0.41%) pooled estimates. Diesel exhaust (*right*) is one source of fine particles at center of debate.

THE TWO-WAY STREET

- Nothing is purely statistical, but most of science and policy has a statistical aspect
- Collaboration is essential
  - We need to understand the subject area
  - Collaborators need to understand statistical science
- Avoid a statistical monopoly; the analogue of:

"Only historians are allowed to reminisce"

THE CLINICIAN'S ROLE

L. Lasagna (1955)

"It is currently fashionable in some circles to consider the clinician member of the team as some sort of minor excrescence, 'a fifth cousin about to be removed.' "

THE STATISTICIAN'S ROLE

A. Bradford Hill (1966)

"Once the clinician has grasped the simple techniques that have been brought to his aid, the statistician has no further part to play. Along with the old soldier he can fade away, contentedly if, sometimes, wistfully."

LEVEL 0 DESIGN

CONSTITUTING INVESTIGATIVE TEAMS

STUDY EFFECTS OF ELF RADIO WAVES

- The Navy considered it an engineering problem
- The investigative team did not have a statistician
  - Study design, conduct and analysis was poor
- The NAS review panel also did not have a statistician!

STROKE TREATMENT WITH t-PA REVIEW PANEL

- NINDS constituted a panel of statisticians to review "statistical" issues
- Mike O'Fallon insisted on including clinicians
  - Issues are statistical & clinical
- Absent clinicians, the panel would be quite ineffective and non-credible

RULES OF ENGAGEMENT

- Insist on proper relations, roles and resources
  - Be "at the table" from the beginning
  - Proactively participate and communicate
- Clarify expectations, rewards and responsibilities
  - Consulting or collaboration?
  - Who does what, when?
  - Administrative controls and roles
- Resources
  - Budgets, staffing and infrastructure
- Authorship based on intellectual contributions

THE RESPECT OF
STATISTICAL SCIENCE
AND STATISTICIANS
BY OTHERS WILL BE
NO GREATER THAN
THE RESPECT
WE CONVEY